MOBILE PRODUCT RECOGNITION WITH EFFICIENT BAG-OF-PHRASE VISUAL SEARCH

Dajiang Zhang, Kim-Hui Yap and Sinduja Subbhuraam

School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore

ABSTRACT

This paper presents a mobile product recognition system using bag-of-visual phrase (BoP). It aims to develop a mobile product recognition and recommendation system where a user can recognize a commercial product of interest by taking a picture of it using the mobile phone, and then search for the relevant information (e.g., price, nearby store, consumer recommendation, etc.). In the proposed BoP framework, second-order visual phrases candidates are first obtained from neighborhood visual words. Discriminative visual phrases are then determined, and images are indexed with a two-dimensional inverted index of visual phrases. Geometric verification (GV) is performed to further improve the accuracy of image matching. Experimental results show that the proposed method can achieve 90% recognition rate for a dataset consisting of 3882 reference images and 41 categories.

Index Terms— Bag-of-Words, Bag-of-Phrases, Mobile Product Recognition

1. INTRODUCTION

As the number of mobile phones has grown tremendously in recent years, more and more mobile applications have been developed for these devices. An emerging application is mobile shopping which has attracted growing attentions. This motivates us to develop a mobile product recognition system that enables users to search product of interest by image queries to find out relevant information (e.g., price, nearby store, consumer recommendation, etc.).

One of the most popular methods for visual search and image matching is the Bag-of-Words (BoW) technique [1-3]. The BoW method learns a visual codebook using K-means clustering based on robust local features (e.g., SIFT [4]) that are extracted from reference images. Each image is represented as a histogram of visual words by quantizing its local descriptors to the nearest centroids. The online search can be efficiently conducted with an inverted index.

Even though the BoW methods do not utilize the spatial cues of local features, it can still provide reasonable performance in some applications (e.g., book cover, CD cover [5]). However, the mobile product recognition poses unique challenges. Compared to other domains such as book or CD covers, commercial products are often 3D objects that span across different shapes, structures or viewpoints. It is also more challenging to distinguish foreground product of interests from the background clutters. Therefore, the spatial correlation of local features is important in the matching procedure. Figure 1 provides some sample images from our test set to demonstrate those issues. These test images are collected by our volunteers.

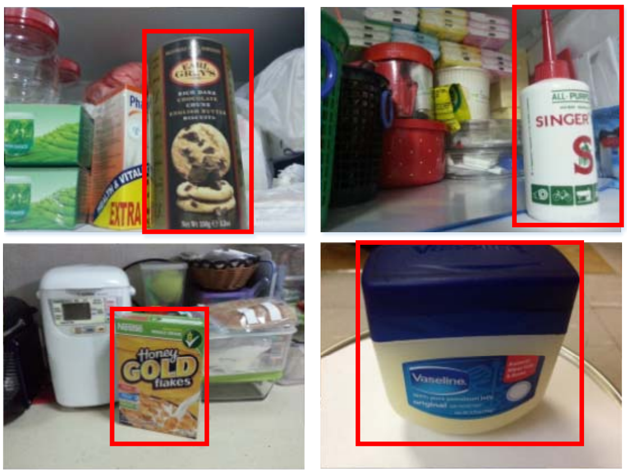

Figure 1. Sample test images collected by volunteers. The products of interest are highlighted by red rectangles.

To utilize the spatial information, combining BoW with geometric verification (GV) can improve the image matching accuracy. However, the online computational cost of GV is expensive, so it can only be performed on a limited number of top ranked candidates. This limits the power of GV because in many cases, the relevant image candidates may not even be short-listed by the BoW method.

Recently, researchers [6-9] have utilized the spatial and contextual information of local descriptors by grouping multiple neighboring visual words into visual phrases. Visual phrases are more robust to cluttered background and have better discriminative power.

In this paper, we propose an efficient BoP framework for mobile product recognition. First of all, candidates of visual phrases are generated by bundling nearby visual words. Then discriminative visual phrases for each category are selected based on their frequency information. Ideally, matches between higher-order visual phrases indicate higher confidence or reliability. But they also suffer from low repeatability. Therefore, we only consider the second-order visual phrases, i.e., visual phrases consisting of two visual words. The online image matching process using visual phrases is conducted efficiently with a two-dimensional inverted index. Geometric verification is then performed on the top-ranked images to further improve the recognition accuracy.

The rest of the paper is organized as follows. Visual phrase generation and selection is introduced in Section 2. Section 3 discusses inverted index and similarity measure with visual phrases. Section 4 provides experimental results and discussions. Section 5 concludes this paper.

2. DISCRIMINATIVE VISUAL PHRASE SELECTION

In the proposed method, SIFT features are extracted for all images. And then a visual vocabulary tree (VT) [10] with depth factor 5 and branch factor 10 is constructed by applying hierarchical K-means clustering to the SIFT descriptors extracted from the reference images. The constructed vocabulary tree is denoted as $V = \{w_1, \ldots, w_N\}$ where N = 100000. The vocabulary tree and the associated inverted index are illustrated in Figure 2.

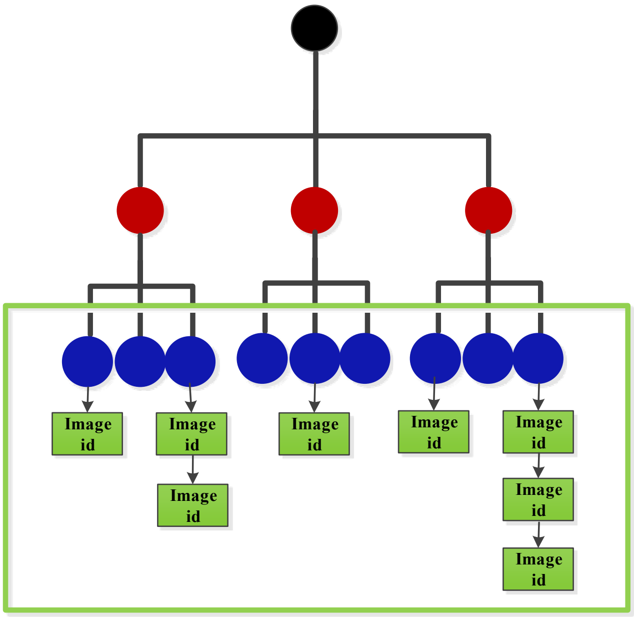

Figure 2. A sample vocabulary tree (with depth factor 2, branch factor 3) and the associated inverted index

Various studies have been done to identify the visual phrases in the literature. In [9], *k* nearest neighbors of local descriptors are grouped as visual phrases. In [8], images are partitioned into fixed-size grids to bundle local features. However, those methods are sensitive to scale changes, i.e., visual phrases extracted from different scales can be very different. In this paper, we utilize the approach in [7] to generate visual phrase candidates. A circular neighboring region is defined around each interest keypoint. The combinations of the central visual word and each of the visual words located in this region form a set of visual phrase candidates. This approach can adjust the radius *r* of the neighboring region with the following equation:

$$r = scale \bullet \lambda \tag{1}$$

The *scale* parameter is the scale where the keypoint is detected [4], and λ is a parameter that determines the range of the neighborhood. Intuitively, λ should be large enough to ensure most useful pairs of visual words will be included. However, large λ will introduce spurious visual phrases and increase the computational cost. In our method, we experimentally set λ as 8 because it is observed that the recognition accuracy no longer improves when λ>8.

The initial pool of visual phrase candidates for each image tends to be noisy, i.e., a large proportion of the visual phrases are unstable and non-representative. In view of this, we perform a two stages procedure to find discriminative visual phrases (DVP). First, those infrequent visual phrases that do not meet a minimum frequency of 2 are eliminated. However, simply removing infrequent visual phrases is not sufficient, since a visual phrase that has a consistently high occurrence across all image categories will not be discriminative. Hence, we further re-rank all visual phrases with the criterion proposed in [7]:

$$Score(i,c) = freq(i,c) / \ln(freq(i,A)) \tag{2}$$

where *freq(i,c)* is the frequency of visual phrase candidate *i* in category *c* and *freq(i,A)* is the frequency of visual phrase *i* in the database. Finally, visual phrases with high *Score(i,c)* are retained.

3. EFFICIENT VISUAL PHRASE INDEXING AND MATCHING

To efficiently compute the image similarity based on BoP representation, we propose to build a two-dimensional inverted index for visual phrases. Since a visual phrase $v = \left\{ w_i, w_j \middle| 1 \le i, j \le N \right\}$ is determined by two visual words, an approach would be to create a matrix of $N \times N$ dimension in which every element represents a possible visual phrase index. The advantage of this structure is that the time complexity of locating a visual phrase index would be $O(1)$. However, the matrix tends to be very sparse and hence it would occupy too much memory. Therefore, we first construct a one-dimension inverted index contains entries for all N visual words in the codebook. These entries correspond to the first visual word w_i in a visual phrase. For the second dimension, we discard null entries that are not pointing to any reference images and sort all the remaining visual word entries in ascending order. To avoid duplicate visual phrase entry, we sort the two visual words of each phrase in ascending order such that $w_i \le w_j$. This inverted index structure is illustrated in Figure 3. The complexity of indexing the first visual word is still $O(1)$. And the complexity of indexing the second one is $O(\log N)$ as we perform a binary search to find the particular entry. Therefore, the online computational complexity of locating a visual phrase entry is $O(\log N)$.

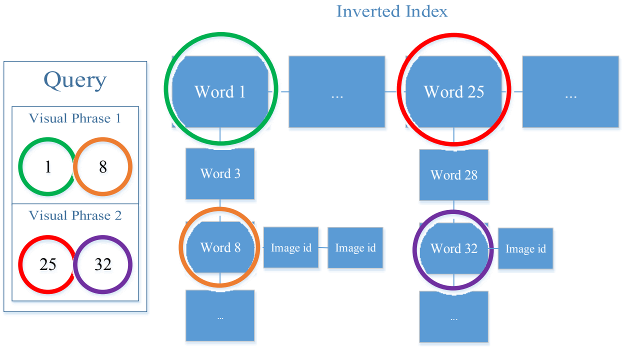

Figure 3. Inverted index of visual phrase and the online querying procedure.

In the proposed framework, histogram intersection is used as the BoP similarity metric. For two histograms H_1 and H_2 of dimension M, the histogram intersection is given as:

$$f(H_1, H_2) = \sum_{n=1}^{M} \min\{h_1^n, h_2^n\} \qquad (3)$$

where h_i^n is the *n-th* element of histogram H_i.

The recognition performance of the BoP is generally superior to the BoW. However, some spurious visual phrases may be obtained from background clutters. Therefore, we further apply a RANSAC-based geometric verification on the top ranked BoP image candidates to improve the recognition accuracy. The GV is conducted by first determining a putative feature matching set $\boldsymbol{C}$ between a pair of images. Then a sample of m matching features is randomly selected from $\boldsymbol{C}$. A model M is computed using coordinates of the selected sample points. Putative matches that are consistent with model M are assumed as inliers. The inlier ratio is computed as $\omega = I / size(C)$. GV will keep iterating the above steps until the probability of finding a better model is less than a predefined value 1-*p*. The GV method is summarized in Algorithm 1.

```
Algorithm 1  Geometric Verification
Input: Putative correspondence set $\boldsymbol{C}$
Start:
  $k := 0, I_{\max} = 0$
  While $k < MaxIters$ do
      $k = k + 1$
      Random select a sample $S$ of size $m$ from $\boldsymbol{C}$
      Compute model $M$ from sample $S$
      Count inlier number $I$ by fitting model $M$ against $\boldsymbol{C}$
      If $I > I_{\max}$ then
          $I_{\max} = I, \omega = I_{\max} / size(C)$
          $MaxIters = \log(1-p) / \log(1-\omega^m)$
      End If
End While
```

4. EXPERIMENTS

4.1. Experimental Setup

To evaluate the performance of the proposed framework, we build an image dataset collected from 41 different commercial products. These images are collected by 10 volunteers. In this dataset, 3882 reference images are acquired from cameras, and 322 test images are captured by mobile phone with various imaging conditions such as different viewpoints, lighting, scales, cluttered backgrounds, etc. All images are resized to a height or width equals to 640 pixels.

In the proposed mobile product recognition system, a query image is considered as correctly recognized if its ground-truth category is the same as the category of the best matched image. The performance is evaluated using recognition rate, which is the percentage of the recognized query images. All experiments are conducted on an Intel 2.0 GHz Quad Core computer with C++ implementation

4.2. Performance Evaluation

The comparison of recognition performance between the baseline BoW and the proposed BoP methods is given in Table 1. It is observed that the proposed BoP method outperforms the BoW method by 11.2%. This is because visual phrases bundle pairs of visual words that are more discriminative than individual visual words.

Table 1. Performance comparison of the Baseline BoW and the proposed BoP methods

Methods	Basedline BoW	Proposed BoP
Recognition Rate (%)	76.7	87.9

Next, we evaluate the performance of using geometric verification to re-rank top-ranked image candidates of the baseline BoW and proposed BoP methods. Figure 4 shows the recognition rates of the baseline BoW+GV and the proposed BoP+GV with top 10, 20, 30, 40 and 50 images. The GV method improves both the recognition performances of the BoW and BoP. However, the proposed BoP+GV method outperforms the baseline BoW+GV in all the experiments.

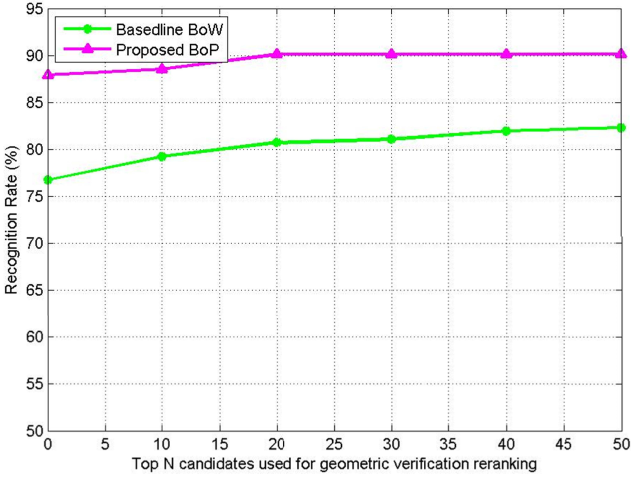

Figure 4. Recognition performance comparison of the proposed BoP and BoW

4.3. Computational Cost

Apart from accuracy, system latency is also an important performance indicator for mobile image recognition application. The average image query time (feature extraction time excluded) for the proposed BoP and baseline BoW methods are given in Figure 5. Note that when there is no GV re-ranking, the proposed BoP method (11.1 milliseconds) and the BoW method (8.6 milliseconds) are both quite efficient. Although the GV increases the running time for both the BoP and Bow methods, it is noticed that the GV performed on BoP is more efficient than GV on BoW. This is because the putative BoP matches between two images contain a larger percentage of inliers, and thus the correct image transformation could be found with fewer number of RANSAC iterations.

5. CONCLUSIONS

This paper proposes an efficient BoP framework for mobile product recognition. We investigate the effectiveness of bundling nearby visual word pairs as discriminative visual phrases. The online image matching is performed with a two-dimensional inverted index. The experimental results show that the proposed BoP+GV method can achieve a 90% recognition rate for a dataset consisting of 3882 reference images and 41 categories

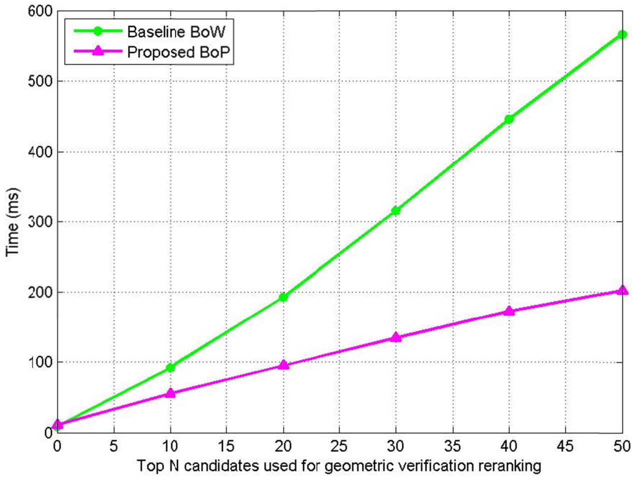

Figure 5. Computational time of the proposed BoP method.

16. REFERENCES

[1] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, pp. 1-8.

[2] K.-H. Yap, T. Chen, Z. Li, and K. Wu, "A comparative study of mobile-based landmark recognition techniques," *Intelligent Systems, IEEE,* vol. 25, pp. 48-57, 2010.

[3] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 1470-1477.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision,* vol. 60, pp. 91-110, 2004.

[5] V. R. Chandrasekhar, D. M. Chen, S. S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, *et al.*, "The stanford mobile visual search data set," in *Proceedings of the second annual ACM conference on Multimedia systems*, 2011, pp. 117-122.

[6] T. Chen, K.-H. Yap, and D. Zhang, "Discriminative bag-of-visual phrase learning for landmark recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 893-896.

[7] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proceedings of the 17th ACM international conference on Multimedia*, 2009, pp. 75-84.

[8] Y. Jiang, J. Meng, and J. Yuan, "Grid-based local feature bundling for efficient object search and localization," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, 2011, pp. 113-116.

[9] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, pp. 1-8.

[10] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 2161-2168.